Computers Math. Applic. Vol. 22, No. 8, pp. 21–34, 1991
Printed in Great Britain.
0097-4943/91 \$3.00 + 0.00

MINIMIZATION OF QUADRATIC FUNCTIONALS ON CONES IN HILBERT SPACES

PAWEŁ J. SZABŁOWSKI
Institute of Mathematics, Technical University of Warsaw
00-661 Warszawa, pl. Politechniki 1, Poland

(Received September 1990)

Abstract—This paper concerns a problem of minimization of a quadratic functional on a cone in a Hilbert space. First, a simple auxiliary minimization in $\mathbf{R}^2$, is solved. The obtained results allow the construction of many different algorithms that solve the primary minimization problem. However, only two algorithms are considered in the paper. The first one concerns a general case, i.e., the case when the only restrictions imposed on the cone are its closeness and convexity. This algorithm uses a projection on the cone technique. The second one can be applied if it is assumed, additionally, that the considered Hilbert space is separable and has Riesz basis. The cone is, in this case, defined as a collection of points whose expansions with respect to this basis have nonnegative coefficients.

Weak convergence of both algorithms is proved. The rates of their convergence are estimated by the convergence of certain series of reals. Possible applications to an optimal control problem are suggested in the introduction and in the conclusion.

1. INTRODUCTION

Although quadratic problems have been considered in the literature [1,2] for many years, there is in fact a lack of algorithms of minimization of a quadratic functional in infinite dimensional spaces. On the other hand, such algorithms can be helpful in solving problems of dynamic optimization.

In 1977, Wierzbicki and Kurcyusz [3] considered a problem of minimizing a functional defined on the Banach space $\mathbf{B}$, with constraints defined by a certain cone in a Hilbert space. Several examples presented there suggested that the considered problem is valid and important. They also presented there the most important properties of the projection on a cone. Besides, they indicated algorithms which could solve the analyzed minimization problem.

All the algorithms mentioned in [3] are based on the notion of the penalty function and the shifted penalty function. Though general, these algorithms have some numerical disadvantages. First of all, they are two stage algorithms; i.e., in order to perform the minimization within the whole set of constraints, one solves infinitely many auxiliary minimization problems. Each of them is a finite dimensional one and special methods of optimization without constraints are chosen. Second, as they are relatively universal, they cannot be the best for a special type of minimized functionals such as quadratic functional.

It is well known that the best algorithms of optimization with constraints (from the numerical point of view) are the Rosen's algorithms (see [4]), which use the projection on the constraints techniques.

The first of the algorithms considered is somewhat similar to the 'gradient projection on the constraints' algorithm in the sense that one of the possible 'directions of improvement' of the functional can be the projection on the cone of the certain variable $-v$ (−gradient) which 'measures the violation of the constraints.' In fact, at each step two possible directions of improvement are considered, one of which is the projection on the cone of $-v$, the second projection of v, and the better of the two directions is chosen.

The second algorithm considers, at each step, all (or in modified, more realistic versions, finite but increasing number of) the elements of the Riesz basis. The best of the possible directions considered is chosen as the actual direction of improvement of the functional.

Moreover, every approximation of the minimization problem satisfies certain condition which will be described below. Suppose that we have the minimization problem:

$$Q(y) = \min_{y \in Y_p} Q(y), \quad Y_p = \{y \in \mathbf{B} : p - P(y) \in \mathbf{C}\},$$

where $\mathbf{B}$ is certain Banach space, $p \in \mathbf{X}$ a Hilbert space with scalar product $\langle \cdot, \cdot \rangle$, $\mathbf{C}$ is a closed, convex cone in $\mathbf{X}$ with vertex in zero, $P : \mathbf{B} \to \mathbf{X}$ an operator and $Q : \mathbf{B} \to \mathbf{R}$ a functional (compare with [3, Section 3]). Further, let $L(\eta, y) = Q(y) + \langle \eta, P(y) - p \rangle$ be a normal Lagrange functional for the minimization problem mentioned above and let $\eta(\hat{y})$ be a normal Lagrange multiplier for the same problem, calculated at the optimal point $\hat{y} \in Y_p$. It is known that both conditions must be satisfied at $\hat{y}$:

$$\eta(\hat{y}) \in \mathbf{C}^* \quad \text{and } \langle \eta(\hat{y}), P(\hat{y}) - p \rangle = 0.$$

($\mathbf{C}^*$—cone dual to $\mathbf{C}$.)

Suppose that we know the functional relationship $\eta(y)$. Each approximation $y^{(k)}$ of the optimal point obtained by the algorithms considered satisfies the following relationship:

$$\langle \eta(y^{(k)}), P(y^{(k)}) - p \rangle = 0.$$

For the quadratic functional, $\mathbf{B} = \mathbf{X}$, $p = 0$ and $P(y) = y$, one can easily find an expression for the mapping $\eta(y)$. η is linear.

One can easily notice that the results of this paper can be extended without any difficulty to all optimization problems for which values of the mapping $\eta(y)$ can be computed explicitly for each $y \in Y_p$.

2. STATEMENT OF THE PROBLEM

Let $\mathbf{X}$ be a Hilbert space with a scalar product $\langle \cdot, \cdot \rangle$ and the norm $\|\cdot\|$, i.e., $\|x\|^2 = \langle x, x \rangle$, $x \in \mathbf{X}$. Let us consider a quadratic functional defined in the following way:

$$m \in \mathbf{X}, \quad M(m) = \langle w, m \rangle + 0.5 \langle m, Hm \rangle,$$

where $w \in \mathbf{X}$, H is a linear, bounded, self adjoint and positive semidefinite operator, with $D(H) = \mathbf{X}$. Suppose that $\mathbf{C} \subset \mathbf{X}$ is a convex closed cone in $\mathbf{X}$. The objective is to find such a point $\mu \in \mathbf{C}$ that:

$$M(\mu) = \inf_{m \in \mathbf{C}} M(m). \tag{1}$$

We will assume the following:

$\mathcal{A}1$ The vector μ defined by (1) exists and is unique.

$\mathcal{A}2$ $\forall m \in \mathbf{C}$, $\langle m, w \rangle < 0 \Longrightarrow \langle m, Hm \rangle > 0$.

$\mathcal{A}3$ A set $\mathbf{X} \supset \mathfrak{M} \overset{\text{def}}{=} \{m \in \mathbf{C}; \langle m, w + Hm \rangle = 0\}$ is bounded in $\mathbf{X}$.

We next present the following basic lemma.

LEMMA 1. *A necessary and sufficient condition for the vector μ to be optimal, i.e., $M(\mu) = \inf_{m \in \mathbf{C}} M(m)$, is*

$$\mu \in \mathbf{C}, \tag{2a}$$
$$\langle \mu, w + H\mu \rangle = 0, \tag{2b}$$
$$w + H\mu \in \mathbf{C}^*, \tag{2c}$$

where $\mathbf{C}^*$ is a closed convex cone dual to $\mathbf{C}$.

For the proof see [3].

As it follows from Lemma 1, the optimal vector μ must satisfy the three conditions (2a), (2b) and (2c). On the other hand, notice that conditions (2a) and (2b) define the set $\mathfrak{M}$. Hence, the optimal vector μ must satisfy the following two conditions: $\mu \in \mathfrak{M}$ and $w + H\mu \in \mathbf{C}^*$.

The idea of the algorithm considered in this paper is to generate, somehow, a sequence of vectors $\{m_i\}$ satisfying the first of the above conditions (i.e., $m_i \in \mathfrak{M}$) and then to try to satisfy the second one 'better and better.' The idea is to measure somehow the violation of the third condition and thus try to minimize this violation.

However, before one considers how to find recursively the sequence $\{m_i\}$, one has to be able to solve the following auxiliary minimization problem:

For a given $m \in \mathfrak{M}$ and $d \in \mathbf{C}$, find two reals t and a, such that

$$tm + ad \in \mathfrak{M} \quad \text{and } M(tm + ad) = \min_{\substack{\alpha,\beta \in \mathbf{R} \\ \alpha m + \beta d \in \mathbf{C}}} M(\alpha m + \beta d). \tag{1'}$$

This problem of minimization is in fact a two dimensional one and it would be easy to solve if, first, the operator H were not positive semidefinite, and second, if $\mathfrak{M}$ were not a subset of $\mathbf{C}$.

We will solve problem (1′) completely using geometrical methods. We will also give several properties of the optimal vector $tm + ad$ (for given $m \in \mathfrak{M}$ and $d \in \mathbf{C}$) (see Lemma 4). These properties will help to formulate the basic properties of the considered algorithms quickly. The following section is important for the considerations of Sections 4 and 5.

3. AUXILIARY MINIMIZATION PROBLEM

Suppose now that we have a vector $m \in \mathfrak{M}$ that is not optimal in the sense that $w + Hm \in \mathbf{C}^*$. Suppose further that we have chosen, somehow, a vector $d \in \mathbf{C}$. (A discussion on how to do this will be given later on.) We will find two reals, t_d and a_d, such that the vector $\bar{m} = t_d\, m + a_d\, d \in \mathbf{C}$ and, moreover, that $M(\bar{m}) < M(m)$. t_d and a_d will be chosen in the best way in the following sense:

$$M(t_d\, m + a_d\, d) = \min_{(t,a) \in E_d \cap G_d} M(tm + ad), \tag{3}$$

where $E_d = \{(t,a)' \in \mathbf{R}^2 : \langle tm + ad,\, w + H(tm + ad)\rangle = 0\}$, $G_d = \{(t,a)' \in \mathbf{R}^2 : tm + ad \in \mathbf{C}\}$. Notice that $E_d \cap G_d = \{(t,a)' \in \mathbf{R}^2 : tm + ad \in \mathfrak{M}\}$.

Let us denote, for simplicity: $J = \langle m, Hm\rangle$, $\gamma(d) = \langle d, Hm\rangle$,

$$A_d = \begin{pmatrix} J & \gamma(d) \\ \gamma(d) & \langle d, Hd\rangle \end{pmatrix}, \quad b_d = \begin{pmatrix} J \\ -\langle w, d\rangle \end{pmatrix}, \quad x_1 = t, \quad x_2 = a.$$

Thus, $E_d = \{x \in \mathbf{R}^2 : x' A_d\, x = b_d'\, x\}$, where $(\cdot)'$ denotes the transposition of $(\cdot)$ ($x' A_d\, x$ is thus a quadratic form and $b_d'\, x$ is a scalar product on $\mathbf{R}^2$).

REMARK 1. *If $d \in \mathbf{C}$ is chosen in such a way that $\langle d, Hd\rangle = 0$, then for $m \in \mathfrak{M}$, we have:*

$$\min_{(t,a) \in E_d \cap G_d} M(tm + ad) = M(m), \qquad (\text{that is, } t_d = 1 \text{ and } a_d = 0).$$

PROOF. $\langle d, Hd\rangle = 0 \Longrightarrow \gamma(d) = 0$, since A_d has to be positive semidefinite. $M(tm + ad) = 0.5\, t^2 J - t\, J + a\langle d, w\rangle$. Now notice that we must have $\langle d, w\rangle \geq 0$. It is so because if $\langle d, w\rangle < 0$, then $\langle d, Hd\rangle > 0$ by $\mathcal{A}2$. Notice also that, since for $(t,a) \in E_d \cap G_d$, $\min M(tm + ad)$ has to be defined uniquely, we must have $a \geq 0$. Thus, $\min_{t,a \geq 0} M(tm + ad) = .5J = M(m)$. So it is enough to take $t = 1$ and $a = 0$. ∎

It follows from the above remark that it does not make sense to consider $d \in \mathbf{C}$ such that $\langle d, Hd\rangle = 0$. Hence, we will consider only such vectors $d \in \mathbf{C}$ with $\langle d, Hd\rangle > 0$. Moreover, for technical reasons, we will assume that the vector $d \in \mathbf{C}$ is selected in such a way that $\gamma(d) + \langle w, d\rangle = \langle d, w + Hm\rangle \neq 0$. The vectors d having this property can be always found, since if $\forall d \in \mathbf{C}$, $\langle w + Hm, d\rangle = 0$, then this would mean that $w + Hm \in \mathbf{C}^*$, that is $m (\in \mathfrak{M})$ would be optimal. Hence, we will assume additionally:

$\mathcal{A}4$ $\langle d, Hd\rangle = 1$, $\gamma(d) + \langle w, d\rangle \neq 0$.

In order to describe the sets G_d and E_d, we will need the following two functionals defined on $\mathbf{C} \times \mathbf{C}$:

$$\Delta(m,d) = \sup\{\alpha \in \mathbf{R} : -\alpha m + d \in \mathbf{C}\}, \quad \sigma(m,d) = \sup\{\alpha \in \mathbf{R} : m - \alpha d \in \mathbf{C}\}, \qquad m, d \in \mathbf{C}.$$

Hence, Δ and σ describe the geometry of the cone $\mathbf{C}$. The properties of these functionals are presented in the following proposition.

PROPOSITION 1.

(i) $\Delta(m,d) \geq 0$, $\sigma(m,d) \geq 0$, *and* $1 - \Delta(m,d)\,\sigma(m,d) \geq 0$, *for* m, $d \in \mathbf{C}$.
(ii) $\forall \lambda \in [0,1]$, $\Delta(m, \lambda\, d_1 + (1-\lambda)\, d_2) \geq \lambda\, \Delta(m,d_1) + (1-\lambda)\, \Delta(m,d_2)$, $\sigma(\lambda\, m_1 + (1-\lambda)\, m_2, d) \geq \lambda\, \sigma(m_1,d) + (1-\lambda)\, \sigma(m_2,d)$, *for all* m, m_1, m_2, d, d_1, $d_2 \in \mathbf{C}$.
(iii) *If* $\Delta(\cdot, d)$ *is nonzero for* m_1 *and* m_2, *then for* $\lambda \in [0,1]$,

$$\frac{1}{\Delta(\lambda\, m_1 + (1-\lambda)\, m_2, d)} \leq \frac{\lambda}{\Delta(m_1,d)} + \frac{(1-\lambda)}{\Delta(m_2,d)}.$$

(iv) *If* $\sigma(m, \cdot)$ *is nonzero for* d_1 *and* d_2, *then for* $\lambda \in [0,1]$,

$$\frac{1}{\sigma(m, \lambda\, d_1 + (1-\lambda)\, d_2)} \leq \frac{\lambda}{\sigma(m,d_1)} + \frac{(1-\lambda)}{\sigma(m,d_2)}.$$

PROOF. (i) follows from relationships $0m + d \in \mathbf{C}$, $m - 0d \in \mathbf{C}$ and the fact that for $\sigma(m,d) \neq 0$, $(-1/\sigma(m,d))\, m + d \notin \mathbf{C}$. (ii), (iii), (iv) follow almost straightforwardly from the closeness and convexity of $\mathbf{C}$ and the definitions of $\Delta(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$. ∎

Using these two functionals, we can easily describe the set G_d. We have:

LEMMA 2. *G_d is a closed, convex cone in $\mathbf{R}^2$ such that:*

$$\begin{aligned} G_d &= \{x \in \mathbf{R}^2 : x_1 \geq -\Delta(m,d)\, x_2\} \cap \{x \in \mathbf{R}^2 : x_1\, \sigma(m,d) \geq -x_2\}, \\ \delta G_d &= R_1(d) \cup R_2(d), \end{aligned} \tag{4}$$

where δA denotes the boundary of A, and

$$\begin{aligned} R_1(d) &= \{x \in \mathbf{R}^2 : -\Delta(m,d)\, x_2 = x_1;\ x_1 \leq 0\}, \\ R_2(d) &= \{x \in \mathbf{R}^2 : \sigma(m,d)\, x_1 = -x_2;\ x_1 \geq 0\}. \end{aligned}$$

PROOF. In the Appendix.

LEMMA 3. *A1–A4* $\Rightarrow (0,0)'$, $(1,0)'$, $(0, -\langle w, d\rangle)' \in E_d$.

PROOF. In the Appendix.

Now notice that, for $x = (t,a)'$, we have $M(tm + ad) = .5\, x' A_d\, x - b_d'\, x$. We have to minimize it on $E_d \cap G_d$. E_d is, however, described by the equation $x' A_d\, x = b_d'\, x$. Hence, on E_d, $M(tm+ad) = -.5 x' A_d\, x$. Thus, we can consider another, equivalent problem of maximization of $L_d(x) = x' A_d\, x$ on $E_d \cap G_d$. Let us denote by $x^d = (t_d, a_d)'$, the point where $L_d(x)$ assumes its maximum on $E_d \cap G_d$.

If $\det A_d = J - \gamma^2(d) > 0$, then $L_d(\cdot)$, considered on E_d, assumes its maximum at

$$x^s = \left(\frac{J + \gamma(d)\langle w, d\rangle}{J - \gamma^2(d)}, \frac{-J(\gamma(d) + \langle w, d\rangle)}{J - \gamma^2(d)} \right), \text{ and}$$
$$L_d(x^s) = J\left(1 + \frac{(\gamma(d) + \langle w, d\rangle)^2}{J - \gamma^2(d)}\right).$$

Hence, if $x^s \in G_d$, then $x^d = x^s$. Otherwise, one would suspect that x^d is equal to either of the two points at which rays $R_1(d)$ and $R_2(d)$ cross E_d. That is, if $x^s \notin G_d$, then the maximum of L_d is reached on the boundary of $E_d \cap \delta G_d$. However, this requires to be proved. Besides, one needs also analytical criteria to judge at which of the boundary points of $E_d \cap G_d$ the maximum is reached, in the case $x^s \notin G_d$. The following theorem will give a complete solution of the problem mentioned above.

Theorem 1. *Let us define:*

$$f_1(d) = (J + \gamma(d)\langle w, d\rangle)\,\sigma(m,d) - J(\gamma(d) + \langle w, d\rangle),$$
$$f_2(d) = (J + \gamma(d)\langle w, d\rangle) - \Delta(m,d)\,J(\gamma(d) + \langle w, d\rangle).$$

(i) *If $f_1(d) \geq 0$ and $f_2(d) \geq 0$, then $J - \gamma^2(d) > 0$. If $\sim(f_1(d) \geq 0$ and $f_2(d) \geq 0)$, then*

$$\gamma(d) + \langle w, d\rangle > 0 \iff f_1(d) < 0,$$
$$\gamma(d) + \langle w, d\rangle < 0 \iff f_2(d) < 0.$$

(ii) $x^d = (t_d, a_d)'$ *where:*

$$t_d = \begin{cases} \dfrac{J + \gamma(d)\langle w, d\rangle}{J - \gamma^2(d)}, & \text{if } f_1(d) \geq 0 \text{ and } f_2(d) \geq 0. \\ \text{Otherwise,} & \\ \qquad \dfrac{J + \sigma(m,d)\langle w, d\rangle}{J - 2\gamma(d)\,\sigma(m,d) + \sigma^2(m,d)} \geq 0, & \text{if } \gamma(d) + \langle w, d\rangle > 0, \\ \qquad -\Delta(m,d)\,a_d \leq 0, & \text{if } \gamma(d) + \langle w, d\rangle < 0. \end{cases}$$

$$a_d = \begin{cases} \dfrac{-J\,(\gamma(d) + \langle w, d\rangle)}{J - \gamma^2(d)}, & \text{if } f_1(d) \geq 0 \text{ and } f_2(d) \geq 0. \\ \text{Otherwise,} & \\ \qquad -\sigma(m,d)\,t_d \leq 0, & \text{if } \gamma(d) + \langle w, d\rangle > 0, \\ \qquad \dfrac{(-J\,\Delta(m,d) - \langle w, d\rangle)}{J\,\Delta^2(m,d) - 2\gamma(d)\,\Delta(m,d) + 1} \geq 0, & \text{if } \gamma(d) + \langle w, d\rangle < 0. \end{cases}$$

$$L_d(x^d) = \begin{cases} J\left(1 + \dfrac{(\gamma(d) + \langle w, d\rangle)^2}{J - \gamma^2(d)}\right), & \text{if } f_1(d) \geq 0 \text{ and } f_2(d) \geq 0. \\ \text{Otherwise,} & \\ \qquad t_d^2\,(J - 2\gamma(d)\,\sigma(m,d) + \sigma^2(m,d)), & \text{if } \gamma(d) + \langle w, d\rangle > 0, \\ \qquad a_d^2\,(J\,\Delta^2(m,d) - 2\gamma(d)\,\Delta(m,d) + 1), & \text{if } \gamma(d) + \langle w, d\rangle < 0. \end{cases}$$

Proof. In the Appendix. ∎

We also have the following lemma, describing the properties of the optimal solution. For a brief formulation of these properties, let us denote: $m_d = t_d\,m + a_d\,d$, $v = w + Hm$, $v_d = w + Hm_d$.

Lemma 4.

(i) $\langle d, v_d\rangle = \langle d, w\rangle + t_d\,\gamma(d) + a_d$

$$= \begin{cases} 0, \ \text{if } f_1(d) \geq 0 \text{ and } f_2(d) \geq 0. & \\ \text{Otherwise,} & \\ \qquad \dfrac{(-f_1(d))}{J - 2\gamma(d)\,\sigma(m,d) + \sigma^2(m,d)} \geq 0, & \text{if } \gamma(d) + \langle w, d\rangle > 0, \\ \qquad \dfrac{(-\Delta(m,d)\,f_2(d))}{J\,\Delta^2(m,d) - 2\gamma(d)\,\Delta(m,d) + 1} \geq 0, & \text{if } \gamma(d) + \langle w, d\rangle < 0. \end{cases}$$

(ii) $\langle m, v_d\rangle = t_d\,J + a_d\,\gamma(d) - J$

$$= \begin{cases} 0, \ \text{if } f_1(d) \geq 0 \text{ and } f_2(d) \geq 0. & \\ \text{Otherwise,} & \\ \qquad \sigma(m,d)\,\langle d, v_d\rangle \geq 0, & \text{if } \gamma(d) + \langle w, d\rangle > 0, \\ \qquad \dfrac{f_2^-(d)}{J\,\Delta^2(m,d) - 2\gamma(d)\,\Delta(m,d) + 1} \geq 0, & \text{if } \gamma(d) + \langle w, d\rangle < 0. \end{cases}$$

Hence,

$$\langle m - \sigma(m,d)\,d, v_d\rangle = \begin{cases} 0, & \text{if } f_2(d) \geq 0, \\ \langle m, v_d\rangle(1-\sigma(m,d)\,\Delta(m,d)), & \text{if } f_2(d) < 0. \end{cases}$$

$$\langle -m\,\Delta(m,d) + d, v_d\rangle = \begin{cases} 0, & \text{if } f_1(d) \geq 0, \\ \langle d, v_d\rangle(1-\sigma(m,d)\,\Delta(m,d)), & \text{if } f_1(d) < 0. \end{cases}$$

(iii) $\langle m,\, H(m_d - m)\rangle = \langle m,\, v_d\rangle$.
(iv) $\langle m_d,\, v\rangle = a_d\,(\gamma(d) + \langle w,\, d\rangle) = J - \langle m_d,\, H\, m_d\rangle + \langle m,\, v_d\rangle < 0$.
(v) $\langle m_d - m,\, H\,(m_d - m)\rangle = \langle m_d,\, H\, m_d\rangle - J - 2\langle m,\, v_d\rangle > 0$.
(vi) $\langle m_d,\, H\, m_d\rangle = t_d\, J - a_d\langle w,\, d\rangle \geq J$.
(vii) $t_d + \gamma(d)\, a_d/J = 1 + \langle m,\, v_d\rangle/J \geq 1$.

PROOF. In the Appendix. ∎

Since we know the value of $L_d(x^d)$, the best choice for d would be the one maximizing $L_d(\cdot)$. To find such a vector is practically impossible. First, the functional L_d has a very complicated form. Second, to find such a vector would mean to be able to solve the minimization problem (1) in one step and get the solution right away. This is, however, impossible. The derived formulae can be useful, however, in examining properties of the optimal vector. Hence, it remains to consider other, unoptimal choices of vector d. Different rules of such choices would lead to different algorithms of minimization of $M(m)$ on the cone $\mathbf{C}$. We will discuss only two such algorithms.

4. ALGORITHMS WITH PROJECTION ON THE CONE

For each $m \in \mathfrak{M}$, find $v(m) = w + Hm$. $v(m)$ does not belong to $\mathbf{C}^*$, because if it did, it would be optimal. It is well known [3] that each $x \in \mathbf{X}$ can be uniquely decomposed in the following way: $x = x^{\mathbf{C}} + x^{-\mathbf{C}^*}$, where x^A denotes projection on A. Let us define

$$\mathbf{C} \ni d_1(m) = (-v(m))^{\mathbf{C}}, \quad \mathbf{C} \ni d_2(m) = (v(m))^{\mathbf{C}}.$$

We will try to 'improve' in the best of the two directions. The second direction is chosen in order to increase the speed with which the 'improper' m decreases. Notice that (following [3, Theorem 2.4]) that:

$$\begin{aligned} \langle d_1(m), w\rangle + \gamma(d_1(m)) &= \langle d_1(m), v(m)\rangle = -||d_1(m)||^2 < 0, \\ \langle d_2(m), w\rangle + \gamma(d_2(m)) &= \langle d_2(m), v(m)\rangle = ||d_2(m)||^2 > 0. \end{aligned}$$

Thus, if we chose d_1 then x^{d_1} lies either in G_{d_1} or on $R_1(d_1)$. On the other hand, if we chose d_2 then x^{d_2} lies either in G_{d_2} or on $R_2(d_2)$. Let us consider the following algorithm:

1. Find the zero approximation $m^{(0)}$ of the optimal vector μ. It can be done, for example, by taking the vector $(-w)^{\mathbf{C}}$ and normalize it by a scalar $b > 0$, in such a way that $b(-w)^{\mathbf{C}} \in \mathfrak{M}$. (Notice that, if $(-w)^{\mathbf{C}} = 0$, then $w \in \mathbf{C}^*$. This, however, would imply that $\min_{m\in\mathbf{C}} M(m) = M(0)$.)
2. Suppose that we have $m^{(k)}$, the k-th approximation of μ. Let us denote $v^{(k)} = w + Hm^{(k)}$, $J_k = \langle m^{(k)}, Hm^{(k)}\rangle$, $\hat{d}_1^{(k)} = (-v^{(k)})^{\mathbf{C}}$, $\hat{d}_2^{(k)} = (v^{(k)})^{\mathbf{C}}$. Denote also:

$$d_i^{(k)} = \begin{cases} \dfrac{\hat{d}_i^{(k)}}{\sqrt{\langle \hat{d}_i^{(k)}, H\,\hat{d}_i^{(k)}\rangle}}, & \text{if } \langle \hat{d}_i^{(k)}, H\,\hat{d}_i^{(k)}\rangle \neq 0, \\ 0, & \text{otherwise,} \end{cases} \qquad \text{for } i = 1,2.$$

Define also the following reals: $\xi_i^{(k)} = \langle w,\, d_i^{(k)}\rangle$, $\gamma_i^{(k)} = \langle d_i^{(k)}, H\, m^{(k)}\rangle$, $\psi_i^{(k)} = \xi_i^{(k)} + \gamma_i^{(k)}$, $\nu_i^{(k)} = \sigma(m^{(k)}, d_i^{(k)})$, $\chi_i^{(k)} = \Delta(m^{(k)}, d_i^{(k)})$, $f_{1i}^{(k)} = (J_k + \gamma_i^{(k)}\,\xi_i^{(k)})\,\nu_i^{(k)} - J_k\,\psi_i^{(k)}$, $f_{2i}^{(k)} = J_k + \gamma_i^{(k)}\,\xi_i^{(k)} - \chi_i^{(k)}\,J_k\,\psi_i^{(k)}$, $i = 1, 2$, $k = 0, 1, 2, \ldots$.

$$\tau_i^{(k)} = \begin{cases} \dfrac{J_k + \xi_i^{(k)}\,\gamma_i^{(k)}}{J_k - (\gamma_i^{(k)})^2}, & \text{if } i = 2 \text{ and } (f_{12}^{(k)} \geq 0), \text{ or } i = 1 \text{ and } (f_{21}^{(k)} \geq 0). \\ \text{Otherwise,} & \\ \qquad \dfrac{J_k + \nu_2^{(k)}\,\xi_2^{(k)}}{J_k - 2\gamma_2^{(k)}\,\nu_2^{(k)} + (\nu_2^{(k)})^2}, & \text{if } i = 2, \\ \qquad \chi_1^{(k)}\,\alpha_1^{(k)}, & \text{if } i = 1. \end{cases}$$

$$\alpha_i^{(k)} = \begin{cases} \dfrac{(-J_k\,\psi_i^{(k)})}{J_k - (\gamma_i^{(k)})^2}, & \text{if } i = 2 \text{ and } (f_{12}^{(k)} \geq 0), \text{ or } i = 1 \text{ and } (f_{21}^{(k)} \geq 0). \\ \text{Otherwise,} & \\ \qquad -\nu_2^{(k)}\,\tau_2^{(k)}, & \text{if } i = 2, \\ \qquad \dfrac{(-J_k\,\psi_1^{(k)} - \xi_1^{(k)})}{J_k\,(\chi_1^{(k)})^2 - 2\gamma_1^{(k)}\,\chi_1^{(k)} + 1}, & \text{if } i = 1. \end{cases}$$

$$F_k(i) = \tau_i^{(k)} - 1 - \frac{\alpha_i^{(k)}\,\xi_i^{(k)}}{J_k}, \qquad i = 1, 2.$$

3. Find an index i_k, such that $F_k(i_k) = \max(F_k(1), F_k(2))$ and define:

$$a_k = \begin{cases} \alpha_1^{(k)}, & \text{if } i_k = 1, \\ \alpha_2^{(k)}, & \text{if } i_k = 2, \end{cases} \qquad t_k = \begin{cases} \tau_1^{(k)}, & \text{if } i_k = 1, \\ \tau_2^{(k)}, & \text{if } i_k = 2. \end{cases}$$

4. Take $m^{(k+1)} = t_k\, m^{(k)} + a_k\, d_{i_k}^{(k)}$.
5. Go to 2.

REMARK 2. Comparing assertions (vi) and (i) of Lemma 4 with the formulae defining $F_k(i)$, $i = 1, 2$, one can easily notice that:

$$F_k(i) = \frac{\langle \tau_i^{(k)}\, m^{(k)} + \alpha_i^{(k)}\, d_i^{(k)},\, H(\tau_i^{(k)}\, m^{(k)} + \alpha_i^{(k)}\, d_i^{(k)})\rangle - J_k}{J_k}. \tag{5}$$

Notice also that using assertions (i) and (ii) of Lemma 4, we have:

$$F_k(i) = \frac{-\langle d_i^{(k)},\, v^{(k)}\rangle\, \alpha_i^{(k)}}{J_k} + \frac{\langle m^{(k)},\, w + H\,(\tau_i^{(k)}\, m^{(k)} + \alpha_i^{(k)}\, d_i^{(k)}\rangle}{J_k}. \tag{6}$$

Thus, $F_k(i_k) = -\langle d_{i_k}^{(k)}, v^{(k)}\rangle a_k / J_k + \langle m^{(k)}, v^{(k+1)}\rangle / J_k$.

Further, using the assertions of Lemma 4, we have:

$$\langle m^{(k)},\, v^{(k+1)}\rangle = \begin{cases} \langle m^{(k)},\, w + H\,(\tau_1^{(k)}\, m^{(k)} + \alpha_1^{(k)}\, d_1^{(k)})\rangle \geq 0, & \text{if } i_k = 1, \\ \langle m^{(k)},\, w + H\,(\tau_2^{(k)}\, m^{(k)} + \alpha_2^{(k)}\, d_2^{(k)})\rangle \geq 0, & \text{if } i_k = 2, \end{cases} \tag{7}$$

$$\langle m^{(k)},\, H\,(m^{(k+1)} - m^{(k)})\rangle = \langle m^{(k)}, v^{(k+1)}\rangle, \tag{8}$$

$$\langle m^{(k+1)},\, v^{(k)}\rangle = J_k - J_{k+1} + \langle m^{(k)},\, v^{(k+1)}\rangle = a_k\, \langle d_{i_k}^{(k)},\, v^{(k)}\rangle \leq 0, \tag{9}$$

$$\langle m^{(k+1)} - m^{(k)},\, H\,(m^{(k+1)} - m^{(k)})\rangle = J_{k+1} - J_k - 2\langle m^{(k)},\, v^{(k+1)}\rangle, \tag{10}$$

$$J_{k+1} = t_k\, J_k - a_k\, \langle w,\, d_{i_k}^{(k)}\rangle. \tag{11}$$

We have also the following theorem.

THEOREM 2.
Assume $\mathcal{A}1$–$\mathcal{A}3$, then

(i) $M(m^{(k)}) \xrightarrow[k\to\infty]{} \inf_{m\in \mathbf{C}} M(m)$.

(ii) $\exists \mu \in \mathbf{C},\ v \in \mathbf{C}^*;\ m^{(k)} \xrightarrow[k\to\infty]{} \mu$ *weakly in* $\mathbf{X}$, $w + H\, m^{(k)} = v^{(k)} \xrightarrow[k\to\infty]{} v$ *weakly in* $\mathbf{X}$.

(iii) $\sum_k \|(-v^{(k)})^{\mathbf{C}}\|^4 < \infty$, $\sum_k \langle m^{(k+1)} - m^{(k)}, H\,(m^{(k+1)} - m^{(k)})\rangle < \infty$; *if* $v \neq 0$, *then* $\sum_k \sigma(m^{(k)}, d_2^{(k)}) < \infty$, $\tau_1^{(k)} \xrightarrow[k\to\infty]{} 1$, $\alpha_1^{(k)} \xrightarrow[k\to\infty]{} 0$, $\langle d_1^{(k)}, w + H\, m^{(k)}\rangle \xrightarrow[k\to\infty]{} 0$.

PROOF. In the Appendix. ∎

5. ALGORITHMS FOR THE MINIMIZATION OF THE QUADRATIC FUNCTIONAL ON A SPECIAL TYPE OF CONES

We will assume now that the space $\mathbf{X}$ is a separable Hilbert space with a Riesz basis $\{\mathbf{e}_i\}_1^\infty$. Suppose that this basis is such that the cone $\mathbf{C} \subset \mathbf{X}$ is defined as follows:

$$\mathbf{C} = \{x \in \mathbf{X} : x = \sum_k \beta_k\, \mathbf{e}_k;\ \beta_k \geq 0,\ k = 1,2,\dots\}.$$

It is easy to notice that:

$$\mathbf{C}^* = \{x \in \mathbf{X} : \langle x, \mathbf{e}_i\rangle \geq 0,\ i = 1,2,\dots\}.$$

It is also easy to calculate the values of the functionals σ and Δ on $\mathbf{C} \times \mathbf{C}$. That is, if we use the convention $g/0 = \infty$ for $0 < g \in \mathbf{R}$, we have for $m = \sum_i \beta_i\, \mathbf{e}_i$, $d = \sum_i \gamma_i\, \mathbf{e}_i$, and then

$$\Delta(m,d) = \sup\{\Delta : -\Delta\, m + d \in \mathbf{C}\} = \sup\{\Delta : \Delta \leq \frac{\gamma_i}{\beta_i}\} = \inf_i \left(\frac{\gamma_i}{\beta_i}\right),$$

$$\sigma(m,d) = \sup\{\sigma : m - \sigma\, d \in \mathbf{C}\} = \sup\{\sigma : \frac{\beta_i}{\gamma_i} \geq \sigma\} = \inf_i \left(\frac{\beta_i}{\gamma_i}\right).$$

Hence, in particular, $\Delta(m, \mathbf{e}_i) = \begin{cases} 1/\beta_i, & \text{if } \beta_i \neq 0 \\ 0, & \text{if } \beta_i = 0 \end{cases}$, $\sigma(m, \mathbf{e}_i) = \beta_i$, for $m = \sum_i \beta_i\, \mathbf{e}_i$. Moreover, one can always find, without any difficulty, a projection on $\mathbf{C}$ of any element of $\mathbf{X}$. If $v = \sum_i \beta_i\, \mathbf{e}_i$, then $v^{\mathbf{C}} = \sum_{i \in I_v} \beta_i\, \mathbf{e}_i$, where $I_v = \{i : \beta_i \geq 0\}$.

Using the properties mentioned above, one can simplify and clarify the algorithm considered in Section 3. This algorithm has, however, a serious disadvantage concerning practical computer applications. Namely, if $\mathbf{X}$ is infinite dimensional, then the computer always uses a finite dimensional approximation of all elements considered in this space. Hence, in order to avoid unpredictable numerical errors, one should modify the algorithm by adjusting it to the practical situation. It is relatively easy to construct such modification. Namely, let us consider a sequence of cones $\{\mathbf{C}_k\}_0^\infty$, such that $\mathbf{C}_{k-1} \subset \mathbf{C}_k \subset \mathbf{C}$, $k = 1,2,\dots$; $\mathbf{C}_k$ is finite dimensional and $\bigcup \mathbf{C}_k = \mathbf{C}$. Find $m^{(0)} \in \mathbf{C}_0$ and further $m^{(k)} \in \mathbf{C}_k$, and choose $d \in \mathbf{C}_{k+1}$. Hence, each $m^{(k)}$ will be finite dimensional and there will be no fixed approximation of the infinite dimensional space by the finite dimensional one. Of course, such a modification requires the space $\mathbf{X}$ to be separable. On the other hand, it will not affect the weak convergence properties of the algorithm. Notice that the above remark concerns also the algorithm considered in the previous section.

We will consider now another algorithm that is conceptually simpler and will allow a more straightforward estimation of the speed of its convergence. Consider the situation described in the introduction and assume $\mathcal{A}1$–$\mathcal{A}3$. Instead of $\mathcal{A}4$, we will assume the following:

$\mathcal{A}4'$ Operator H and the basis $\{\mathbf{e}_i\}$ are such that $\langle \mathbf{e}_i, H\,\mathbf{e}_i\rangle = 1$, $i = 1,2,\dots$ and, moreover, we will consider only such $\mathbf{e}_i$, for which $\langle \mathbf{e}_i, H\, m\rangle + \langle \mathbf{e}_i, w\rangle \neq 0$ for a given m.

Let us denote $\omega_i = \langle w, e_i\rangle$, $i = 1,2,\ldots$. Notice that $\exists i_0$, such that $\omega_{i_0} < 0$. This is true because we have $0 = \langle m, w + H\,m\rangle = \langle m, H\,m\rangle + \sum_i \beta_i\,\omega_i$, where $m = \sum_i \beta_i\,e_i$. Thus, $\sum_i \beta_i\,\omega_i = -\langle m, H\,m\rangle < 0$ for $m \in \mathbf{C}$. Since $\beta_i \geq 0$, we deduce that, indeed, $\exists i_0$, $\omega_{i_0} < 0$. Now, choose an increasing sequence $\{n_k\}$ of integers and consider the following algorithm of minimization:

(1) Find $m^{(0)}$, 0-approximation of the optimal vector μ such that $m^{(0)} \in \mathfrak{M}$ and, moreover, that $J_0 = \langle m^{(0)}, H\,m^{(0)}\rangle \geq (\inf_i \omega_i)^2$. (One can take, for example, $m^{(0)} = -\omega_{i_0}\,e_{i_0}$, if there exists i_0 such that $\omega_{i_0} = \inf_i \omega_i$.)

(2) Suppose that we have $m^{(k)}$, the k-th ($k \geq 0$) approximation of the optimal vector. Define the following auxiliary quantities: $J_k = \langle m^{(k)}, H\,m^{(k)}\rangle$, $\gamma_i^{(k)} = \langle e_i, H\,m^{(k)}\rangle$, $\nu_i^{(k)} = \sigma(m^{(k)}, e_i)$, (notice that now, $m^{(k)} = \sum_i \nu_i^{(k)}\,e_i$), $f_i^{(k)} = (J_k + \gamma_i^{(k)}\,\omega_i)\,\nu_i^{(k)} - J_k\,(\gamma_i^{(k)} + \omega_i)$ for $k = 0,1,\ldots$, $i = 1,\ldots,n_k$.

Consider also the following function $F_k : \{1,\ldots,n_k\} \longrightarrow \mathbf{R}^+$,

$$F_k(i) = \begin{cases} \dfrac{(\gamma_i^{(k)} + \omega_i)^2}{J_k - (\gamma_i^{(k)})^2}, & \text{if } f_i^{(k)} \geq 0, \\[2ex] \nu_i^{(k)}\,\dfrac{(-f_i^{(k)} + (\gamma_i^{(k)} + \omega_i)\,(J_k + \nu_i^{(k)}\,\omega_i))}{J_k\,(J_k - 2\gamma_i^{(k)}\,\nu_i^{(k)} + (\nu_i^{(k)})^2)}, & \text{if } f_i^{(k)} < 0, \end{cases} \qquad i = 1,2,\ldots,n_k.$$

(3) Find i such that $F(i) = \max_{1 \leq i \leq n_k} F_k(i)$ and denote, for simplicity, $\gamma_k^* = \gamma_{i_k}^{(k)}$, $\omega_k^* = \omega_{i_k}$, $\psi_k = f_{i_k}^{(k)}$, $\nu_k^* = \nu_{i_k}^{(k)}$:

$$t_k = \begin{cases} \dfrac{J_k + \gamma_k^*\,\omega_k^*}{J_k - (\gamma_k^*)^2}, & \text{if } \psi_k \geq 0, \\[2ex] \dfrac{J_k + \nu_k^*\,\omega_k^*}{(J_k - 2\gamma_k^*\,\nu_k^* + (\nu_k^*)^2)}, & \text{if } \psi_k < 0, \end{cases}$$

$$a_k = \begin{cases} \dfrac{(-J_k\,(\gamma_k^* + \omega_k^*))}{J_k - (\gamma_k^*)^2}, & \text{if } \psi_k \geq 0, \\[2ex] -\nu_k^*\,t_k, & \text{if } \psi_k < 0. \end{cases}$$

(4) Take $m^{(k+1)} = t_k\,m^{(k)} + a_k\,e_{i_k}$.

(5) Go to (2).

REMARK 3. Notice that, if $J_0 > (\inf_i \omega_i)^2$ then $\forall k$, $J_k > (\inf_i \omega_i)^2$. Now if $J_k > \omega_i^2$ for some i and k, then x^{J_k} has the first coordinate positive and, consequently, $x^{e_i} \in \{x^s\} \cup R_2(e_i)$. Let us consider now the case $J_k < \omega_j^2$ for some j and k. Notice first that then $\omega_j > 0$ since $J_k > (\inf_i \omega_i)^2$. Second, notice that we have also $\gamma_j + \omega_j > 0$, for if it were otherwise, we would have $0 < \omega_j < -\gamma_j \leq \sqrt{J_k}$ and, consequently, $J_k > \omega_j^2$, which contradicts the assumption that $J_k < \omega_j^2$. Hence, if $J_k < \omega_j^2$, then x^{J_k} must lie in the third quadrant (of $\mathbf{R}^2$) and, thus, must have the second coordinate negative. Hence, in either case, we have $x^{e_i} \in \{x^s\} \cup R_2(e_i)$. The algorithm is, thus, well constructed according to the rules derived in Section 3 (e.g., $F_k(i) = t_k - 1 - a_k\,\omega_i/J_k$, and so on).

Using Lemma 4 we get immediately the following properties of the algorithm; let us denote for simplicity: $v^{(k)} = w + H\,m^{(k)}$,

$$\langle e_{i_k}, v^{(k+1)}\rangle = \omega_k^* + t_k\,\gamma_k^* + a_k = \frac{\psi_k}{(J_k - \gamma_k^*\,\nu_k^* + (\nu_k^*)^2)} \geq 0, \tag{12}$$

$$\langle m^{(k)}, v^{(k+1)}\rangle = \nu_k^*\,\langle e_{i_k}, v^{(k+1)}\rangle \geq 0, \tag{13}$$

$$\langle m^{(k)}, H\,(m^{(k+1)} - m^{(k)})\rangle = \langle m^{(k)}, v^{(k+1)}\rangle, \tag{14}$$

$$\langle m^{(k+1)}, v^{(k)}\rangle = a_k\,(\gamma_k^* + \omega_k^*) = J_k - J_{k+1} + \langle m^{(k)}, v^{(k+1)}\rangle < 0, \tag{15}$$

$$\langle m^{(k+1)} - m^{(k)}, H\,(m^{(k+1)} - m^{(k)})\rangle = J_{k+1} - J_k - 2\,\langle m^{(k)}, v^{(k+1)}\rangle, \tag{16}$$

$$J_{k+1} = t_k\,J_k - a_k\,\omega_k^*, \tag{17}$$

$$t_k + \frac{\gamma_k^*\,a_k}{J_k} = 1 + \frac{\langle m^{(k)}, v^{(k+1)}\rangle}{J_k} \geq 1. \tag{18}$$

Moreover, we have the following theorem:

THEOREM 3. *Assume A1–A3, then:*

(i) $J_k \xrightarrow[k\to\infty]{} -2 \min_{m\in\mathbf{C}} M(m)$.

(ii) $\exists \mu \in \mathbf{C}$, $v \in \mathbf{C}^*$, $v^{(k)} \xrightarrow[k\to\infty]{} v$ *weakly in* $\mathbf{X}$, $m^{(k)} \xrightarrow[k\to\infty]{} \mu$ *weakly in* $\mathbf{X}$.

(iii) *Denote by* $I_v = \{i : \langle \mathrm{e}_i, v\rangle > 0\}$, *then* $\forall i \in I_v$, $\sum_k \nu_i^{(k)} < \infty$.

(iv) $\nu_{i_k}^{(k+1)} = 0$, *if* $\psi_k < 0$.

(v) $\forall k$, $i_k \neq i_{k+1}$.

PROOF. In the Appendix. ∎

6. NUMERICAL EXAMPLES

The algorithm considered in Section 5 has been tested by computing some numerical examples in finite dimensional spaces. A finite dimensional example has been chosen purposely since there exist many different algorithms concerning the finite dimensional case. On the other hand, a simple analysis of how the algorithm acts leads to the supposition that the algorithm should work worst in the following situation.

Suppose μ is the optimal vector and i is such an index that $\langle \mathrm{e}_i, \mu\rangle = 0$. Suppose further that for some k we have $\langle m^{(k)}, \mathrm{e}_i\rangle \gg 0$. The algorithm should diminish this value to zero. The supposition is that takes the most of the algorithm's working time.

It was rather difficult to find an example which would be of high dimension. We have found such examples in $\mathbf{R}^5$ and $\mathbf{R}^3$. Thus, $\mathbf{X}$ was either $\mathbf{R}^5$ or $\mathbf{R}^3$. $\{\mathrm{e}_i\}$ was a standard orthonormal basis and the cone is $\mathbf{C} = \{x \in \mathbf{X} : \sum_i \alpha_i \mathrm{e}_i;\ \alpha \geq 0\}$. The operator H was a symmetric positive semidefinite matrix with all its diagonal elements equal to 1. Let $\mathbf{X} = \mathbf{R}^5$. The matrix H was constructed as follows.

First take the 3×5 matrix

$$A = \begin{vmatrix} 1 & 2 & -2 & -2 & -1 \\ 2 & 1 & 2 & 2 & 0 \\ 2 & 2 & -1 & 1 & 0 \end{vmatrix}$$

and generate the 5×5 matrix $\hat{H} = A'A$. Notice that $\operatorname{rank}(\hat{H}) = 3$. Now we normalize $\hat{H}$ in the following way: If $\hat{H} = [\hat{h}_{ij}]$, then let us calculate $h_{ij} = \hat{h}_{ij}/\sqrt{(\hat{h}_{ii}\,\hat{h}_{jj})}$, i, $j = 1, \ldots, 5$. Define $H = [h_{ij}]$. The vector w was defined as $w = [-12.8, -8, -25/3, -12, 7/(3\sqrt{2})]'$. Notice that $w_1 = \min_{1\leq i\leq 5} w_i$. We have taken $m^{(0)} = -w_1\,\mathrm{e}_1$. On the other hand, it turns out that $\mu_1 = 0$. Hence, the example is good.

The second example was similar but concerned $\mathbf{R}^3$, with

$$H = \begin{vmatrix} 1 & .8 & 0 \\ .8 & 1 & -.6 \\ 0 & -.6 & 1 \end{vmatrix}, \text{ and}$$

$w = [-1.8, -1, -1]'$. We took again $m^{(0)} = 1.8\,\mathrm{e}_1$, but $\mu_1 = 0$.

We will not present the details. The general remark that follows from the previous examples (and also from the 'more regular ones') is that the algorithm is characterized by a very quick convergence of the sequence $\{J_k\}$. Practically, after few iterations (2–7, depending on the dimension) we had $(\lim J_k - J_k)/(\lim J_k) \cong .05\%$. The required accuracy (assumed very good, nearly 10^{-8}) was reached after either 13 iterations (in the case of $\mathbf{R}^3$), or after 120 iterations (in the case of $\mathbf{R}^5$). Such 'behaviour' of the algorithm can be explained by the fact that both examples (but especially the one with $\mathbf{R}^5$) concerned the situation when the minimum was very 'flat,' in the following sense. The set

$$\mathbf{A}_\varepsilon = \{m \in \mathbf{C} : \langle m,\, w + H\,m\rangle = 0;\ \langle m, H\,m\rangle < \varepsilon + \lim J_k,\ \varepsilon > 0\}$$

is very 'big' and contains many vectors with $m_1 \gg 0$ and small norm of $(-v)^{\mathbf{C}}$, $v = w + H\,m$.

7. POSSIBLE APPLICATIONS AND CONCLUSIONS

Although the examples presented in [3] concern the more general problem of minimization (recalled in Section 1), they can, however, be easily adjusted to the minimization problem considered here, by taking the functional Q as quadratic and setting $\mathbf{B} = \mathbf{X}$, $p = 0$, $P(y) = y$. Hence, possible applications to the optimal control problems of the algorithms derived in this paper could be easily obtained.

On the other hand, one can think of applying the algorithms presented in the paper to the following general optimization problem with constraints. Suppose that a functional to be minimized is defined on a Hilbert space. Suppose also that the minimized functional is regular enough (say, convex and having Frèchet derivatives up to second order) so that one could approximate it by a quadratic one, at least within the neighborhood of the actual approximation of the optimal point. Further, assume that the constraints could be estimated by a closed, convex cone, also within the neighborhood of the actual approximation. Then one would be able to solve the general problem of minimization of a nonquadratic functional on a general set of constraints, by solving a sequence of simple problems of minimization of quadratic functionals on cones. Thus, one would have a two stage algorithm. If one, however, wanted to apply other methods (such as penalty functional or shifted penalty functional methods) to solve a sequence of problems of minimization of quadratic functional on the cones, then one would get three stage, hence very inaccurate, methods.

Thus, by applying the straightforward algorithms presented in the paper, one is getting simpler, more accurate algorithms of optimization. It should be also emphasized that the idea of substituting a general optimization problem, by the sequence of simpler problems of minimization of quadratic functional on cones, is becoming more and more popular at least in the concerned of literature.

REFERENCES

1. K.C. Kiwiel, A method for solving certain quadratic programming problems arising in nonsmooth optimization, *IMA J. Num. Anal.* **6**, 137–152 (1986).
2. M.J.D. Powell, On the quadratic programming algorithm of Goldfarb and Idnani, *Math. Progr. Study* **25**, 46–61 (1985).
3. A.P. Wierzbicki and St. Kurcyusz, Projection on a cone, penalty functionals and duality theory for problems with inequality constraints in Hilbert space, *SIAM J. Control Optimization* **15** (1), 25–56 (1977).
4. J.B. Rosen, The gradient projection method for nonlinear programming: II. Non-linear constraints, *J. Soc. Indust. Appl. Math.* **9**, 514–532 (1961).
5. J. Musielak, *Wstep do analizy funkcjonalnej*, (in Polish), PWN Warszawa, (1976).

APPENDIX

PROOF OF LEMMA 2. It can be easily seen that G_d is a convex closed cone in $\mathbf{R}^2$, since $\mathbf{C}$ is a closed convex cone in $\mathbf{X}$.

Take $G_d \ni x = (x_1, x_2)'$. Notice that either sign $(x_1 x_2) = -1$ or $x_1 \geq 0$ and $x_2 \geq 0$. If the later case is true then of course $x_1 \geq -\Delta(m,d)\, x_2$ and $x_1\, \sigma(m,d) \geq -x_2$. Thus, consider the case $x_1 \geq 0$ and $x_2 < 0$. We have then $x_1\, m + x_2\, d \in \mathbf{C}$ and, further, $m - (-x_2/x_1)\, d \in \mathbf{C}$. Hence, $-x_2/x_1 \leq \sigma(m,d)$. We argue in a similar way when $x_1 < 0$ and $x_2 \geq 0$. If $x_1 = 0$ then $x_2 > 0$, and if $x_2 = 0$ then $x_1 > 0$. ∎

PROOF OF LEMMA 3. If $x = (0,0)'$ then $x' A_d\, x = 0$ and $b_d'\, x = 0$. If $x = (1,0)$ then $x' A_d\, x = J$ and $x'\, b_d = J$. If $x = (0, -\langle w,d\rangle)'$, then $x' A_d\, x = (\langle w,d\rangle)^2$ and $x'\, b_d = (\langle w,d\rangle)^2$. ∎

In order to prove Theorem 1, we will need the following auxiliary lemmas:

LEMMA A1. *For every $z \in (0, \sup_{x \in E_d} L_d(x))$ there exist two points $x_{1z}, x_{2z} \in E_d$ such that $L_d(x_{1z}) = L_d(x_{2z}) = z$. Moreover, versor $x_{1z} - x_{2z}$ does not depend on z.*

PROOF. By a straightforward calculation we obtain

$$x_{1z} = \left(\frac{z}{J} + \frac{\langle w,d\rangle\, y_1}{J}, y_1\right); \quad x_{2z} = \left(\frac{z}{J} + \frac{\langle w,d\rangle\, y_2}{J}, y_2\right),$$

where:

$$y_1 = \frac{(-(\gamma + \langle w,d\rangle)\, z + \sqrt{\delta})}{J - \gamma^2 + (\gamma + \langle w,d\rangle)^2}, \qquad y_2 = \frac{(-(\gamma + \langle w,d\rangle)\, z - \sqrt{\delta})}{J - \gamma^2 + (\gamma + \langle w,d\rangle)^2}, \text{ and}$$

$$\delta = z\,(J\,(\gamma + \langle w,d\rangle)^2 - (z - J)\,(J - \gamma^2)).$$

If $f_1(d) \geq 0$ and $f_2(d) \geq 0$, then

$$\xi = \langle w, d\rangle + \frac{J\,\gamma(d) + \gamma^2(d)\,\langle w, d\rangle - J\,\gamma(d) - J\,\langle w, d\rangle}{J - \gamma^2(d)} = 0.$$

If $f_1(d) < 0$; $\gamma(d) + \langle w, d\rangle > 0$, then:

$$\xi = \langle w, d\rangle + t_d\,(\gamma(d) - \sigma(m, d)) = \frac{(-f_1(d))}{J - 2\gamma(d)\,\sigma(m, d) + \sigma^2(m, d)}.$$

If $f_2(d) < 0$; $\gamma(d) + \langle w, d\rangle < 0$, then:

$$\xi = \langle w, d\rangle + a_d\,(\Delta(m, d) + 1) = \frac{(-\Delta(m, d))\,(-f_2(d))}{J\,\Delta^2(m, d) + 2\gamma(d)\,\Delta(m, d) + 1}.$$

(ii) Let $\eta = \langle m, v_d\rangle = -J + t_d\,J + a_d + a_d\,\gamma(d)$.
If $f_1(d) \geq 0$ and $f_2(d) \geq 0$, then:

$$\eta = \frac{J^2 + J\,\gamma(d)\,\langle w, d\rangle - J\,\gamma^2(d) - J\,\gamma(d)\,\langle w, d\rangle}{J - \gamma^2(d)} - J = 0.$$

If $\gamma(d) + \langle w, d\rangle > 0$; $f_1(d) < 0$, then:

$$\eta = t_d\,(J - \sigma(m, d)\,\gamma(d)) - J = \frac{\sigma(m, d)\,(-f_1(d))}{J - 2\gamma(d)\,\sigma(m, d) + \sigma^2(m, d)}.$$

If $\gamma(d) + \langle w, d\rangle < 0$; $f_2(d) < 0$, then:

$$\eta = a_d\,(J\,\Delta(m, d) + 1) - J = \frac{(-f_2(d))}{J\,\Delta^2(m, d) + 2\gamma(d)\,\Delta(m, d) + 1}.$$

(iii) $\langle m, H(m_d - m)\rangle = \langle m, v_d - v\rangle = \langle m, v_d\rangle$, since $\langle m, v\rangle = 0$, for $m \in \mathfrak{M}$.
(vi) Since $(t_d, a_d) \in E_d$, we have: $L_d(x^d) = J\,t_d - a_d\,\langle w, d\rangle = b_d'\,x^d$.
(iv) $\langle m_d, v\rangle + \langle m_d, H\,m_d\rangle - J = a_d\,\langle d, v\rangle + t_d\,J - \langle w, d\rangle\,a_d - J = a_d\,\langle w, d\rangle + a_d\,\gamma(d) + J\,t_d - \langle w, d\rangle - J = J\,t_d - J + a_d\,\gamma(d) = \langle m, v_d\rangle$ (compare the proof of ii)).
(v) $\langle m_d - m, H\,(m_d - m)\rangle = \langle m_d, H\,(m_d - m)\rangle - \langle m, H\,(m_d - m)\rangle = \langle m_d, v_d\rangle - \langle m_d, v\rangle - \langle m, v_d\rangle = \langle m_d, H\,m_d\rangle - J - 2\langle m, v_d\rangle$.
(vii) Compare the proof of (ii). ∎

PROOF OF THEOREM 2.

(i) The sequence $\{J_k\}$ is monotone (compare with the construction of the algorithm) and bounded from below (assumption $\mathcal{A}1$), hence, it is convergent.

(ii) It follows from (i) that $F_k(i_k) \xrightarrow[k\to\infty]{} 0$, and consequently, $F_k(i) \xrightarrow[k\to\infty]{} 0$, $i = 1, 2$. Notice also that $\sim (t_k^2 + a_k^2 \xrightarrow[k\to\infty]{} 0)$. The fact that $F_k(1) \xrightarrow[k\to\infty]{} 0$ and (6) imply that $\alpha_1^{(k)}\langle d_1^{(k)}, v^{(k)}\rangle \xrightarrow[k\to\infty]{} 0$. If there existed a subsequence $\{k_i\}$ such that $\alpha_1^{(k_i)} \xrightarrow[i\to\infty]{} 0$, then consequently, we would have $\tau_1^{(k_i)} \xrightarrow[i\to\infty]{} 1$ and $\exists i_0$, $\forall i > i_0$, $\tau_1^{(k_i)} > 0$. This would, however, imply that $f_{21}^{(k)} > 0$ for $i > i_0$.

We would have then

$$\alpha_1^{(k_i)} = \frac{J_{k_i}\,\psi_1^{(k_i)}}{J_{k_i} - (\gamma_1^{(k_i)})^2} = \frac{(-J_{k_i}\,\langle d_1^{(k_i)}, v^{(k_i)}\rangle)}{J_k - (\gamma_1^{(k_i)})^2}.$$

Hence, $\langle d_1^{(k_i)}, v^{(k_i)}\rangle = \|(-v^{(k_i)})^{\mathrm{C}}\|^2 \xrightarrow[i\to\infty]{} 0$. If the sequence $\{k_i\}$ did not exist, then of course $\langle d_1^{(k_i)}, v^{(k_i)}\rangle = \|(-v^{(k_i)})^{\mathrm{C}}\|^2 \xrightarrow[i\to\infty]{} 0$.

Now take any weakly convergent subsequence $\{m^{(k_i)}\}$ of the sequence $\{m^{(k)}\}$. Such subsequence always exists since $m^{(k)} \in \mathfrak{M}$ and $\mathfrak{M}$ is bounded, hence, weakly compact. Let $m^{(k_i)} \xrightarrow[i\to\infty]{} m$. The sequence $v^{(k_i)} = w + H\,m^{(k_i)}$ also weakly converges to such a point v that $(-v)^{\mathrm{C}} = 0$. This means, however, that $v \in \mathrm{C}^*$. As the optimal point μ is uniquely defined, we deduce that $m = \mu$. Thus, we have shown that any weakly convergent subsequence of $\{m^{(k)}\}$ converges to μ. It means, however, that $m^{(k)} \xrightarrow[i\to\infty]{} \mu$ weakly in $\mathbf{X}$. Consequently $v = w + H\,\mu$. Thus, we have shown (i) and (ii).

(iii) Now, since $\langle d_1^{(k)}, v^{(k)}\rangle \xrightarrow[k\to\infty]{} 0$, we deduce that the sequence: $J_k+\gamma_1^{(k)}\xi_1^{(k)} = J_k-(\gamma_1^{(k)})^2+\gamma_1^{(k)}\langle d_1^{(k)}, v^{(k)}\rangle$ tends to some nonnegative limit as $k \longrightarrow \infty$. However,

$$\frac{J_k+\gamma_1^{(k)}\xi_1^{(k)}}{J_k-(\gamma_1^{(k)})^2+(\gamma_1^{(k)}+\xi_1^{(k)})^2}$$

is the first coordinate of the vector x^{J_k}. Hence, the ray $R_s(d_1^{(k)})$ that connects the points (0,0) and $(x^{J_k}+(1,0))/2$ must lie inside $G_{d_1^{(k)}}$ for sufficiently large k. This means that $\exists k_0, \forall k \geq k_0, f_{21}^{(k)} \geq 0$, and consequently, $\tau_1^{(k)} \xrightarrow[k\to\infty]{} 1$ and $\alpha_1^{(k)} \xrightarrow[k\to\infty]{} 0$. On the other hand, for $k \geq k_0$,

$$\frac{\alpha_1^{(k)}\langle d_1^{(k)}, v^{(k_s)}\rangle}{J_k} = \frac{\langle d_1^{(k)}, v^{(k)}\rangle^2}{J_k-(\gamma_1^{(k)})^2} = \frac{\|(-v^{(k)})^{\mathrm{C}}\|^4}{J_k-(\gamma_1^{(k)})^2}.$$

Since $F_k(i_k) = (J_{k+1}-J_k)/J_k$ and J_k converges monotonically, we deduce that $\sum_k F_k(i_k) < \infty$. Consequently, we have $\sum_k F_k(i) < \infty$, $i = 1, 2$. Now it is already easy to deduce that $\sum_k \|(-v^{(k)})^{\mathrm{C}}\|^4 < \infty$.

Similarly, since we have (7), we deduce that $\sum_k \langle m^{(k)}, v^{(k+1)}\rangle < \infty$. Applying (10), we get

$$\sum_k \langle m^{(k+1)} - m^{(k)}, H(m^{(k+1)} - m^{(k)})\rangle < \infty.$$

Further, since $\|(-v^{(k)})^{\mathrm{C}}\| \xrightarrow[k\to\infty]{} 0$, we must have $\liminf_{k\to\infty} \langle v^{(k)}, d_2^{(k)}\rangle > 0$. Since we have also $\sum_k F_k(2) < \infty$ and $F_k(2) > \alpha_2^{(k)}\langle d_2^{(k)}, v^{(k)}\rangle > 0$, we deduce that $\sum_k \left(-\alpha_2^{(k)}\right) < \infty$. Now if there existed a subsequence k_s such that $f_{12}^{(k_s)} \geq 0$, then repeating the argument used already in the proof of (ii), we would have: $\langle v^{(k_s)}, d_2^{(k_s)}\rangle \xrightarrow[s\to\infty]{} 0$.

Hence, $\exists k_1$, such that $\forall k > k_1$, $f_{12}^{(k)} < 0$. This means that: $\alpha_2^{(k)} = \sigma(m^{(k)}, d_2^{(k)})\,\tau_2^{(k)}$, for $k > k_1$. Since $\alpha_2^{(k)} \xrightarrow[k\to\infty]{} 0$ implies $\tau_2^{(k)} \xrightarrow[k\to\infty]{} 1$, we have thus shown that $\sum_k \left(-\alpha_2^{(k)}\right) < \infty$ implies that $\sum_k \sigma(m^{(k)}, d_2^{(k)}) < \infty$. ∎

PROOF OF THEOREM 3.

(i), (ii) We argue similarly as in the proof of the Theorem 2. The sequence $\{J_k\}$ is convergent. Take any weakly convergent subsequence $\{m^{(k_s)}\}$. Let $m^{(k_s)} \xrightarrow[s\to\infty]{} m$. The sequence $v^{(k_s)} = w + H\,m^{(k_s)}$ is also weakly convergent. $F_k(i_k) \xrightarrow[k\to\infty]{} 0$, $F_k(i_k) \geq F_k(i)$, $i = 1, \dots, n_k$. Thus if $n_k \xrightarrow[k\to\infty]{} \infty$, then $\forall i$, $F_k(i) \xrightarrow[k\to\infty]{} 0$.

Further, notice that if for some i we had: $\langle \mathrm{e}_i, v^{(k)}\rangle = \omega_i + \gamma_i^{(k)} < 0$, then

$$F_k(i) = \frac{(\omega_i+\gamma_i^{(k)})^2}{J_k-(\gamma_i^{(k)})^2},$$

hence, if there exists a subsequence $\{n_j\}$ such that $\langle \mathrm{e}_i, v^{(n_j)}\rangle < 0$, $j = 1, 2, \dots$, then $\lim_{j\to\infty} \langle \mathrm{e}_i, v^{(n_j)}\rangle = 0$. Thus, for each fixed i we have $\langle \mathrm{e}_i, v^{(n_j)}\rangle \xrightarrow[j\to\infty]{}$ const. ≥ 0.

Thus, we have shown that $v^{(n_j)} \xrightarrow[j\to\infty]{} v = w + H\,m \in \mathrm{C}^*$. Since μ is defined uniquely, and as $m \in \mathfrak{M}$, then $\langle m, w + H\,m\rangle = 0$. Hence, $m = \mu$. Thus each weakly convergent subsequence $\{m^{(k)}\}$ also converges to μ. This means weak convergence of $\{m^{(k)}\}$ to μ in $\mathbf{X}$.

(iii) We use assertion (iii) of Theorem 2 with $\mathrm{C} = \{x : x = \beta\,\mathrm{e}_i, \beta \geq 0\}$, $i \in I_v$. Then, of course, $(v^{(k)})^{\mathrm{C}}/\|(v^{(k)})^{\mathrm{C}}\| = \mathrm{e}_i = d_2^{(k)}$. Since also for $i \in I_v$, $\exists k_1$, $\varepsilon > 0$, $\langle \mathrm{e}_i, v^{(k)}\rangle > \varepsilon$, $k \geq k_1$, we can apply the same argument as in the proof of the assertion (iii) of Theorem 2 and get $\sum_k \sigma(m^{(k)}, d_2^{(k)}) = \sum_k \sigma(m^{(k)}, \mathrm{e}_i) = \sum_k v_i^{(k)} < \infty$.

(iv) If $\psi_k < 0$, then $\nu_{i_k}^{(k+1)} = t_k\,\nu_{i_k}^{(k)} + a_k = t_k\,\nu_{i_k}^{(k)} - t_k\,\nu_{i_k}^{(k)} = 0$.

(v) First, notice that $\langle \mathrm{e}_{i_k}, v^{(k+1)}\rangle$ is either positive or equal to zero and $\nu_{i_k}^{(k+1)}$ is equal to zero or positive and, moreover, $\nu_{i_k}^{(k+1)}\langle \mathrm{e}_{i_k}, v^{(k+1)}\rangle = 0$. On the other hand, if $\nu_i^{(k)} = 0$ and $\gamma_i^{(k)} + \omega_i \geq 0$ for some i, then $f_i^{(k)} < 0$ and, consequently, $F_k(i) = 0$. Thus, i_{k+1} cannot be equal to i_k unless $m^{(k+1)}$ is optimal. ∎